D880–D884 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1216

H2DB: a heritability database across multiple species by annotating trait-associated genomic loci

Eli Kaminuma[1], Takatomo Fujisawa[1], Yasuhiro Tanizawa[1], Naoko Sakamoto[1], Nori Kurata[2], Tokurou Shimizu[3] and Yasukazu Nakamura[1,*]

[1]Genome Informatics Laboratory, [2]Plant Genetics Laboratory, National Institute of Genetics, 1111 Yata, Mishima 411-8540 and [3]Citrus Research Division, NARO Institute of Fruit Tree Science, Shimizu 424-0292, Japan

Received August 15, 2012; Revised October 16, 2012; Accepted October 31, 2012

ABSTRACT

H2DB (http://tga.nig.ac.jp/h2db/), an annotation database of genetic heritability estimates for humans and other species, has been developed as a knowledge database to connect trait-associated genomic loci. Heritability estimates have been investigated for individual species, particularly in human twin studies and plant/animal breeding studies. However, there appears to be no comprehensive heritability database for both humans and other species. Here, we introduce an annotation database for genetic heritabilities of various species that was annotated by manually curating online public resources in PUBMED abstracts and journal contents. The proposed heritability database contains attribute information for trait descriptions, experimental conditions, trait-associated genomic loci and broad- and narrow-sense heritability specifications. Annotated trait-associated genomic loci, for which most are single-nucleotide polymorphisms derived from genome-wide association studies, may be valuable resources for experimental scientists. In addition, we assigned phenotype ontologies to the annotated traits for the purposes of discussing heritability distributions based on phenotypic classifications.

INTRODUCTION

The cost of DNA sequencing using next-generation sequencers has declined. Thus, trait association studies that require numerous samples and genome-wide genetic markers have now become popular. Trait-associated genomic loci, such as single-nucleotide polymorphism (SNP), genes and quantitative trait loci (QTL), are now summarized as databases in several published studies [genome-wide association studiessnp (GWASsnp) (1), GWASdb (2), SNPedia (3), AnimalQTLdb (4), etc.]. It is important to evaluate the genetic effects of particular genomic loci. In particular, genetic and environmental effects on phenotypic variances are indispensable for basic genetic studies when variances can be biologically defined by statistics that reflect genetic heritability.

Many studies have estimated genetic heritability under various experimental conditions. In general, heritability estimates can be confirmed by results published in research journals. For example, one annotation database of heritability estimates for human traits was extracted from biomedical research journals (3). However, this database is annotated only for human heritability estimates. Thus, there is no collection of heritability estimates among various species other than that for specific organisms.

Therefore, we constructed an annotation database of heritability estimates for heterogeneous organisms from publicly available online journals by manual curation. Information on attributes for extracted heritability was also annotated. Attributes included trait descriptions, experimental conditions, trait-associated genomic loci, different labels for broad- and narrow-sense heritabilities, estimated standard deviations and annotated document information. We also attempted to classify heritabilities by making assignments from phenotypic quality ontology (PATO) (5). PATO provides an ontology of qualities that is commonly described phenotypes with qualitative character in any organism. Thus, it enabled us to classify and compare fitness and non-fitness traits. The proposed heritability database should aid users to view tendencies for heritability estimates across multiple species.

*To whom correspondence should be addressed. Tel: +81 559816859; Fax: +81 559816889; Email: yn@nig.ac.jp

MATERIALS AND METHODS

Data collection

The procedure to annotate quantitative heritability estimates was divided into three stages: (i) annotating heritability estimates and attribute information from digital text documents, (ii) concurrently extracting information of causal SNPs/genes and (iii) assigning PATO ontologies to curated traits. In the first stage, we extracted heritability estimates by examining digital texts selected using the search keyword 'heritability' from NCBI PUBMED abstracts (6). If descriptions in the literature were insufficient, the contents of the online journals were investigated instead of PUBMED abstracts. It must be noted that the results of abstract-based curation do not always cover those of contents-based curation. Heritability estimates were extracted only if the trait values covered a range without extraction, such as when only the average value was determined. Regarding descriptions, standard deviations of heritability estimates and distinctions between broad- and narrow-sense heritabilities were collected. Mandatory annotated items were heritability estimates (0.00–1.00), accession numbers and publication years of the annotated documents, species names, trait descriptions, experimental conditions including population names, standard deviations of heritability estimates and trait-associated genomic loci. Cases of non-significant heritability estimates were not extracted. The database site presents the floating points of heritability estimates rounded to two digits after the decimal point. The annotation data that were collected by literature curation and assigned PATO term with accession number were registered in the postgresql database.

PATO assignments and classifications

For ontology assignments using PATO, we focused on several terms in 'physical object quality' for convenience. In the first layer of PATO, six fields were selected: cellular quality, molecular quality, morphology, organismal quality, physical quality and population quality. A conventional research study (7) that investigated narrow-sense heritabilities classified phenotypic traits into four categories: morphological traits, physiological traits, behavioural traits and life history traits. We could roughly map morphological traits to morphology (PATO:0000051) and its child terms, and life history traits to both reproductive quality (PATO:0001434) and viability (PATO:0000169) in organismal quality. Likewise, we could map a behavioural trait to behavioural quality (PATO:0000186) in organismal quality, and physiological traits to child terms, except for several morphometric traits (e.g. weight), in physical quality (PATO:0001018). All statistical tests for annotated heritability estimates were performed using MATLAB R2010a (The MathWorks Inc., Natick, MA, USA).

RESULTS

The implemented relational database of annotated heritability estimates, H2DB, is freely accessible at http://tga.nig.ac.jp/h2db/. The three functions of H2DB are as follows: data visualization of annotated heritabilities and additional information with annotation statistics, search selection boxes for heritability by species name and PATO term and graphical visualization of frequency distributions of heritability estimates. Figure 1 is a representative screenshot of H2DB. Data visualization is confirmed in the centre table by paging up and down. The upper left part has two selection boxes for searching. The upper right panel shows a histogram of selected heritabilities.

Annotation statistics for heritability estimates and trait-associated genomic loci

Until August 2012, the number of annotated heritability estimates was 838 and the number of unique annotated documents was 189. Document accession numbers in the original database and the publication years can be confirmed. The abbreviation 'A' or 'C' indicates an annotated document from either abstract or journal contents. The annotation type 'contents' included 58% of all heritability entries. The number of entries annotated from the contents tended to be greater than those from the abstracts because document contents were often inspected if a range of heritability estimates that indicated multiple entries was reported in an abstract text. Subsequently, the number of unique annotated species was 67 when crossed lines were independently counted.

Supplementary Table S1 shows the number of entries for all annotated species. The number of entries for humans is 225 (27%). Except for humans, entries for breeding animals and crop plants tended to be most often representative. The two attributes of traits and experimental conditions were annotated independently. Standard deviations of heritability estimates were not always provided, so the extracted number was 264 (32%). The type of heritability estimate was characterized as broad- or narrow-sense when the respective phrases were described in the literature. In the current version, the number of annotated heritability entries for broad- and narrow-sense is 195 (23%) and 115 (14%), respectively. The remaining entries (63%) were unspecified. The number of unique PATO terms is 35 in the current version of the proposed database. Of the 838 heritability entries, 659 entries were assigned to one of the 35 PATO terms. Supplementary Table S2 shows the assigned PATO terms. The total number of annotated trait-associated loci was 44, for which the number of unique documents for all annotated loci was 17. In the future, document contents should be examined further because the frequency of locus names appearing in abstract texts was exceedingly rare, even for major controlling loci.

Search query selection by species name and PATO category

In the upper left part of the top screen in Figure 1, a selection box enumerates the annotated species names as a search query entry, thereby allowing the extraction of heritability annotations for target species only. As a convenient interface, a mouse click on a control button

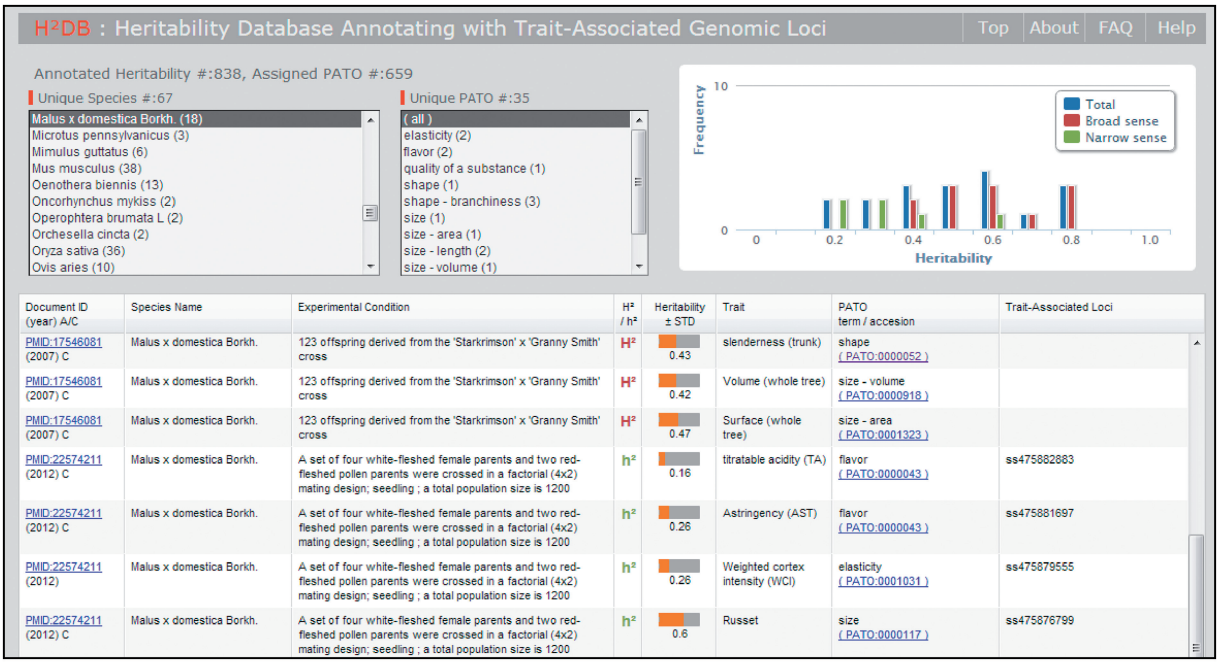

Figure 1. A screenshot of H2DB. Search based on the two selection boxes at the upper left provides annotated heritability that can be narrowed by selected species name and PATO term. Upper right figure is a histogram of selected heritability estimates. In the lower area, a table provides detailed annotated information on heritability.

supports multiple choices for species entries. The number in parenthesis after a species name indicates the number of annotated heritability estimates. A selection box for PATO terms is located next to the species selection box. This search is performed in the same manner as for the selection box for species names. The assignment of PATO terms to classify heritability traits allows users to inspect heritability estimates by classification. Search queries using both species names and PATO terms are restricted to existing entries that have both.

Histogram to investigate categorical distributions of annotated heritabilities

The upper right panel in Figure 1 displays a histogram of quantitatively annotated heritability estimates. This histogram represents total annotated, broad- and narrow-sense heritabilities, where the three types can be distinguished by the bar colours. This histogram allows users to investigate heritability distributions for a specific category using a search function. The following example is an extended analysis for total, broad- and narrow-sense heritabilities related to multiple organisms, where a conventional study (7) focused only on narrow-sense heritability estimated from 75 animal species.

A frequency distribution comprising all 838 annotated heritability estimates was confirmed to be normally distributed ($P < 0.001$) by the Jarque–Bera test. The mean (0.40) and the standard deviation (0.25) of the estimated normal distribution were determined using weighted maximum likelihood estimates. The means and standard deviations for broad- and narrow-sense heritability estimates were 0.51 ± 0.28 and 0.32 ± 0.27, respectively. Figure 2 shows the mean heritabilities by four categories of heritability traits: morphological traits, life history traits, physiological traits and behavioural traits. Percentages in the figure represent the rates of additive genetic effects based on the annotated broad- and narrow-sense heritabilities. The order of narrow-sense heritabilities for four categories was consistent with a conventional study (7). The proposed annotation data incorporating broad-sense heritability generated a new suggestion that the additive genetic effect to the total phenotypic variance may be high for morphological traits as compared with life history (fitness) traits.

However, the amount of data used was too small as shown in Figure 2. The suggestion will be statistically analysed in accumulated data. Then, caution must be taken into consideration while interpreting the results as they were analysed using heterogenous heritability estimates for varying experimental conditions. Several heritability studies reported both broad- and narrow-sense heritabilities concurrently. Thus, future data acquisition for the proposed database will provide more accurate discussions.

FUTURE DIRECTIONS

We have proposed an annotation database of genetic trait heritabilities not only for humans but also other organisms based on online sources of journal abstracts and contents. The proposed database contains attribute information for trait descriptions, experimental conditions, trait-associated genomic loci, standard deviations of heritability estimates and distinctions between broad- and

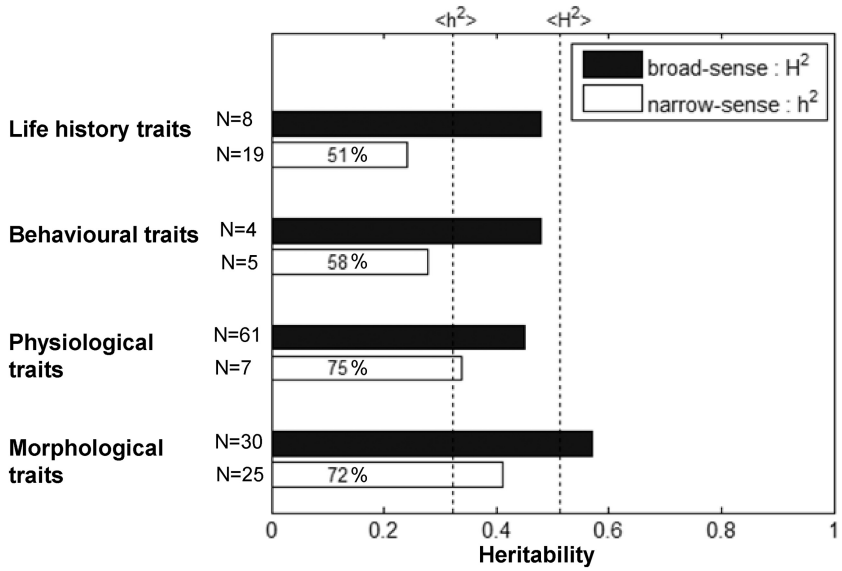

Figure 2. Statistical means of broad- and narrow-sense heritability estimates by four trait categories based on PATO terms (August 2012). The dotted lines indicate the means using all samples. The percentages in the narrow-sense labelled bars indicate the rates of additive genetic effects against the total phenotypic variances (h^2/H^2). *N* = sample size of annotated entries.

narrow-sense heritabilities. We also assigned phenotypic ontologies for the annotated heritabilities for the purposes of classification.

Regarding the development of the proposed database, an important future task will be to consider new definitions of heritability estimates: the missing heritability (8), which is referred to as a discrepancy between narrow-sense heritability and GWAS-driven additive effect, gene–gene interaction effect (9) and gene-by-environment interactions and variance controlling effects (10). To decode characteristic effects of heritabilities may be realized by classification of experimental conditions.

Another significant future task will be to annotate quantitative contribution rates by trait-associated loci against heritability. However, the descriptions provided in online journals are often too vague to specify whether the rate of variance explained by an individual locus was its percentage of the total phenotypic variance or the summed variances of estimated loci. Thus, this will require further investigation for annotating the rates of explained locus-based variances.

For technical developmental, PATO assignments provide only qualitative classifications of annotated traits. To specify a phenotypic trait, entity ontology is required in addition to PATO (5). For entity classifications, organism-specific anatomical ontologies need to be annotated. As an example from plant genetics research, let us consider the quality 'length' (PATO:0000122). When two annotated traits that are assigned to PATO:0000122 include descriptions of leaf length and seed length, plant ontology (11) can provide entity terms: leaf (PO:0025034) and seed (PO:0009010). As shown in (12), annotating entity ontologies for traits would lead to clearer searches of heritability.

Finally, this database system will also be developed to make appropriate connections with variation databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2.

ACKNOWLEDGEMENTS

We gratefully acknowledge Takako Mochizuki, Dr Hideki Nagasaki and Dr Kenta Sumiyama for discussions on database design. In particular, we thank Prof. Toshihiko Shiroishi and Prof. Naruya Saito for their advice on research tools for population genetics.

FUNDING

Transdisciplinary Research Integration Center Project of Research Organization of information and Systems; Japan Society for the Promotion of Science (JSPS) KAKENHI [24500366]. Funding for open access charge: JSPS.

Conflict of interest statement. None declared.

REFERENCES

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.N., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
2. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.
3. Cariaso,M. and Lennon,G. (2012) SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.*, **40**, D1308–D1312.

4. Hu,Z.L., Fritz,E.R. and Reecy,J.M. (2007) AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Res.*, **35**, D604–D609.
5. Gkoutos,G.V., Green,E.C., Mallon,A.M., Hancock,J.M. and Davidson,D. (2005) Using ontologies to describe mouse phenotype. *Genome Biol.*, **6**, R8.
6. Sayers,E.W. *et al.* (2012) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **40**, D13–D25.
7. Mousseau,T.A. and Roff,D.A. (1987) Natural selection and heritability of fitness components. *Heredity*, **59**, 181–197.
8. McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
9. Zuk,O., Hechter,E., Sunyaev,S.R. and Lander,E.S. (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA*, **109**, 1193–1198.
10. Shen,X., Pattersson,M., Ronnegard,L. and Calborg,O. (2012) Inheritance beyond plain heritability: variance-controlling genes in *Arabidopsis Thaliana*. *PLoS Genet.*, **8**, e1002839.
11. Jaiswal,P. *et al.* (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics*, **6**, 388–397.
12. Mungall,C.J., Gkoutos,G.V., Smith,C.L., Haendel,M.A., Lewis,S.E. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.